24.7 An 8.8GHz 198mW 16x64b 1R/1W Variation-Tolerant Register File in 65nm CMOS

Steven Hsu, Amit Agarwal, Mark Anders, Sanu Mathew, Ram Krishnamurthy, Shekhar Borkar

Intel, Hillsboro, OR

Wide bit-width (≥64b) single read/write ported register files are essential building blocks in high-performance superscalar microprocessors demanding single-cycle latency/throughput and dense organization. Significant PVT-induced variations in transistor leakage limit the performance and robustness of wide dynamic register-file BLs. A 16×64b single-read/write ported variation-tolerant register file with single-cycle read/write latency and throughput is fabricated in 65nm CMOS technology [1]. Operating at 8.8GHz, it consumes 198mW (measured at 1.2V, 50˚C) (Fig. 24.7.7). Fused static decode and 64b dynamic array read within a single cycle, split decoder with PVT/burn-in tolerant keeper compensation, leakage-tolerant split WL architecture, and shared twin memory cell (TMC) topology enable a dense layout occupying 0.017mm^2 while simultaneously achieving (i) wide range of PVT operating points across slow-fast corners, 0.5 to 1.4V and 25 to 110˚C, (ii) low active leakage power of 25mW with optimal non-minimum channel-length usage, (iii) high noise immunity with a low BL noise droop ≤8mV, (iv) scalable register file performance up to 10.1GHz, 273mW measured at 1.4V, 50˚C, and (v) low-voltage mode performance of 300MHz, 1.3mW measured at 500mV, 50˚C. Simultaneous supply/temperature scaling enables 13% total power reduction at iso-frequency (measured at 8.8GHz, 1.16V, 25˚C) or 9% higher performance at iso-power (measured at 9.6GHz, 1.25V, 25˚C).

Figure 24.7.1 shows the organization of the single-cycle 1-read, 1-write ported register file tile, consisting of 16-entries × 64b. This design can be scaled to a larger register-file memory array by replication [2]. A 2ϕ 50% duty-cycle clocking plan allows seamless time-borrowing at the address input and data output interfaces. In the first phase (ϕ1) of the cycle, a fully static one-hot 4b address decoder generates the read/write select WLs. Clocked NAND gates convert the static decoder outputs into domino-compatible select WLs in the second phase (ϕ2) of the cycle. The ϕ2 clock is locally generated by inverting the incoming ϕ1 clock, enabling auto-stretchable clocks for slow frequency debug. Four adjacent NAND WL drivers share a common NMOS clock transistor, enabling 36% WL clock power improvement with no performance penalty. Each dynamic local BL (LBL) supports a single-ended read with 8 cells/BL followed by a 2-way merge via static NAND. Figure 24.7.2(a) shows the register file TMC with matched pass transistors on each side of the storage cell, enabling single-ended write with optimal cell stability. This organization enables simultaneous address decodes and 64b read/write operations to non-conflicting locations in a single clock cycle.

The pre-charged 8-way dynamic LBL is susceptible to noise due to high active leakage during evaluate operations. A split-decoder architecture utilizes the most significant bit of the read address to enable/disable the clock drivers of the upper/lower 8×64b array, reducing the pre-charge switching power by 50% (Fig. 24.7.2(b)). Partially decoded address bits permit pre-charge select WLs to transition early, enabling time-borrowing into the dynamic LBL without delay overhead. The strong pre-charge PMOS device anchors the dynamic LBL to the supply, enabling a 2.8× reduction in DC noise droop on unaccessed LBLs.

A split-WL architecture distributes the WL driver inverter across the entire memory array, suppressing input voltage noise offset on the sensitive dynamic pull-down inputs (Fig. 24.7.2(c)). This noise-suppression inverter, locally embedded in the TMC, enables a 35% increase in LBL DC noise robustness at the same performance. WL noise-suppression inverters and wire tracks are shared between 2 bitcells to minimize total area, providing a dense TMC layout occupying 5.09µmx 1.625µm (Fig. 24.7.2(a)). Optimal non-minimum channel lengths [3] in the storage cell cross-coupled inverters and read-access transistors reduce TMC active leakage by 3.7X compared to the conventional bitcell implementation.

A NAND2-based PVT/burn-in keeper with 1b enable provides PVT compensation on all BLs across slow-fast corners, 0.5 to 1.4V, and 25 to 110˚C [4]. Across this PVT range, transistor leakage varies by 4 orders of magnitude. Compared to a 2PMOS-based keeper [5], this implementation enables a 39% reduction in total compensation circuit transistor width (1µm LBL pulldown width, 8% keeper) at equivalent performance (Fig. 24.7.3(a)). With compensation turned on, optimized keeper sizes improve the DC noise robustness of fast dies by 27% to meet the target noise-margin constraints, resulting in a worst-case LBL dynamic node droop ≤8mV. Turning off the compensation on the slow dies trades surplus robustness for higher performance (8% down to 4% keeper). This optimization on slow dies achieves a measured 10% delay improvement and tighter delay/robustness distributions across slow-fast dies (Fig. 24.7.3(b)).

Figure 24.7.4 shows the total single-cycle critical path of 8 gate stages through the fully static decoder and dynamic array bounded by master-slave FFs at the clock boundaries. The register file operates at a maximum frequency (F_{max}) of 8.8GHz (measured at nominal 1.2V, 50˚C), and consumes 198mW total worst-case power with simultaneous 64b read/write operations. Total active leakage power component is 25mW (13% of total power). Figure 24.7.5 shows F_{max}, total power and active leakage power measurements versus supply voltage at 50˚C. Register-file performance is scalable up to 10.1GHz consuming 273mW with an active leakage component of 57mW (measured at 1.4V, 50˚C). In low-voltage mode (measured at 500mV, 50˚C), the register file operates at 300MHz consuming 1.3mW with an active leakage power of 405µW. Figure 24.7.6 shows total power and F_{max} measurements with simultaneous supply/temperature scaling. At 8.8GHz (iso-frequency) operation, supply/temperature scaling to 1.16V, 25˚C reduces total power to 172mW (13% reduction). At 1.25V, 25˚C operation, F_{max} improves to 9.6GHz (9% higher) while maintaining the same total power. PVT compensation is turned on at elevated temperature (110˚C) to enable fully functional operation at iso-frequency (measured at 1.4V, 8.8GHz, 285mW) or iso-power consumption (measured at 1.24V, 7.9GHz, 198mW).

Acknowledgments:
The authors thank the Pyramid Probe Division of Cascade Microtech, Inc. for high-bandwidth wafer level membrane probing solution; C. Webb, N. Saxena, C. Mark, M. Shah, D. Somasekhar, D. Finan for discussions; K. Ikeda and H. Nguyen for layout help; and M. Haycock for encouragement and support.

References:
[1] P. Bai et al., "A 65nm Logic Technology Featuring 35nm Gate Lengths, Enhanced Channel Strain, 8 Cu Interconnect Layers, low-k ILD and 0.57µm^2 SRAM Cell," *IEDM Technical Digest*, pp. 657-660, Dec., 2004.
[2] J. Miller et al., "A 16GB/s, 0.18µm Cache Tile for Integrated L2 Caches from 256kB to 2MB," *Symp. on VLSI Circuits*, pp. 228-231, June, 2000.
[3] E. Fetzer et al., "The Multi-threaded, Parity-Protected 128-word Register Files on a Dual-Core Itanium-Family Processor," *ISSCC Dig. of Tech. Papers*, pp. 382-383, Feb., 2005.
[4] A. Alvandpour et al., "A Burn-in Tolerant Dynamic Circuit Technique," *IEEE Proc. CICC*, pp. 81-84, Sept., 2002.
[5] D. Stasiak et al., "A 2nd Generation 440ps SOI 64b Adder," *ISSCC Dig. of Tech. Papers*, pp. 288-289, Feb., 2000.

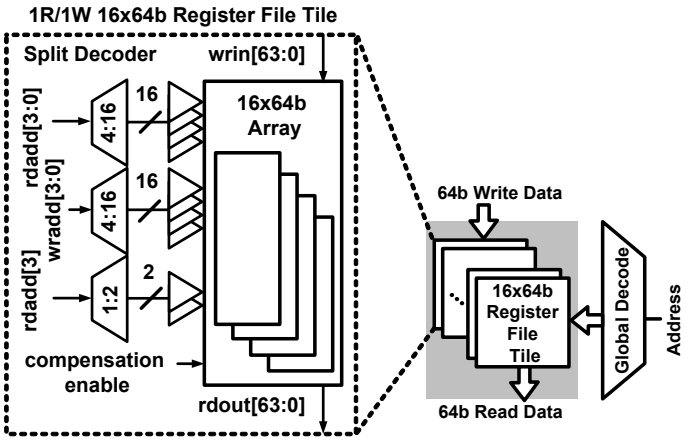


Figure 24.7.1: 1R/1W 16x64b register-file organization.

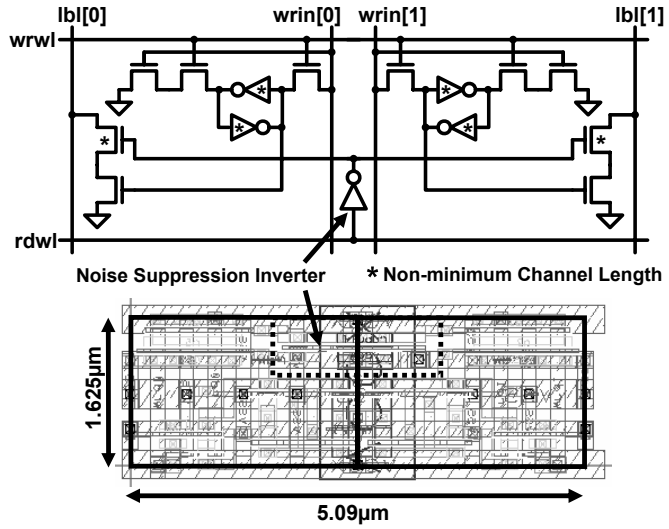


Figure 24.7.2(a): Twin memory cell topology and layout.

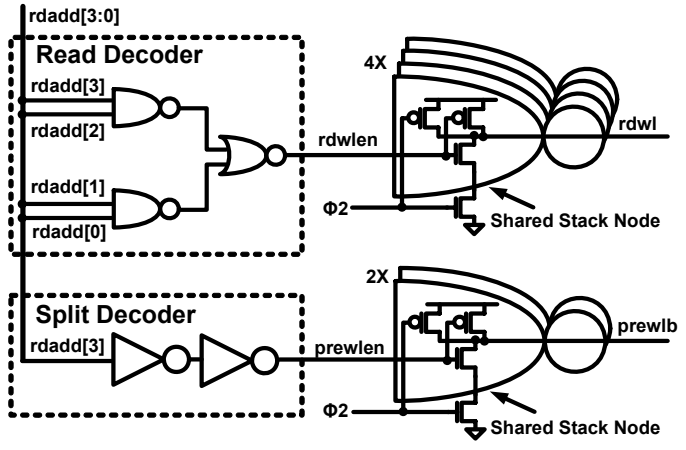


Figure 24.7.2(b): Decoder organization with shared stack node WL drivers.

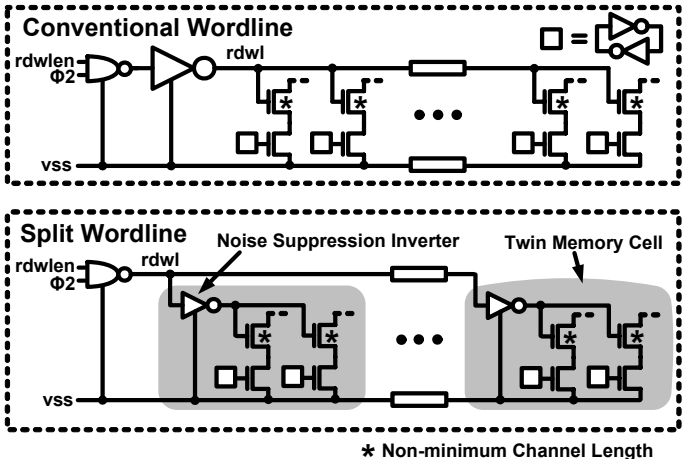


Figure 24.7.2(c): Conventional and split WL circuit topology with noise-suppression inverters.

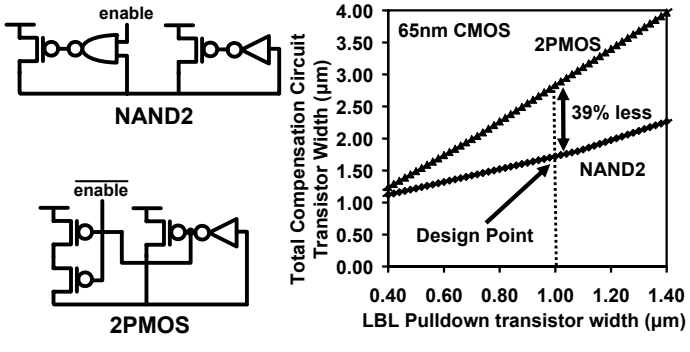


Figure 24.7.3(a): Keeper compensation circuits and comparison.

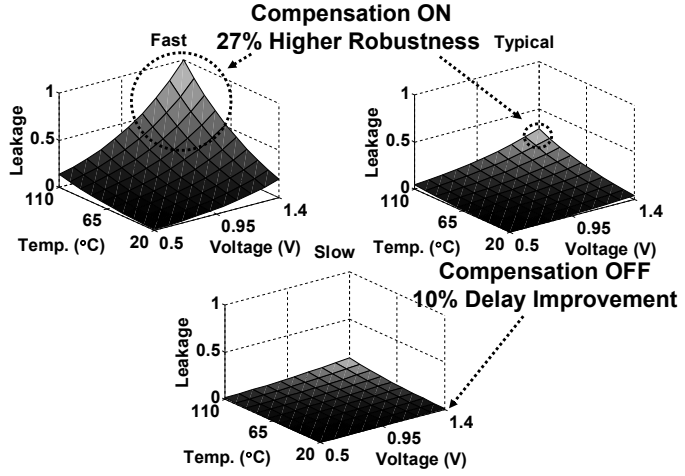


Figure 24.7.3(b): PVT leakage compensation across slow-fast dies.

Continued on Page 664

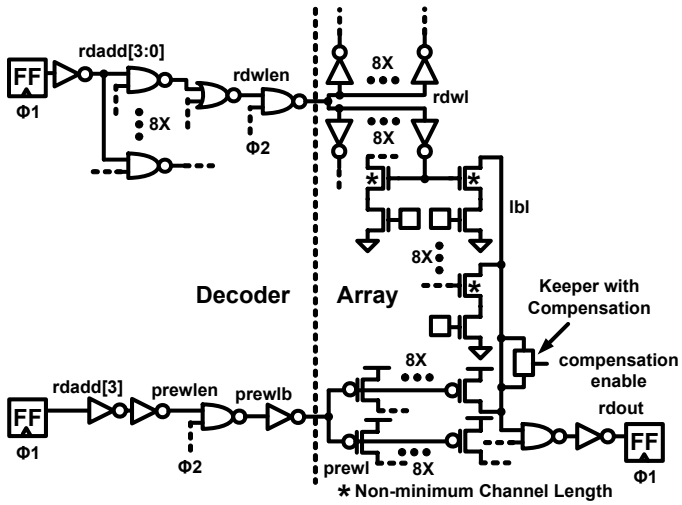


Figure 24.7.4: Register-file tile critical path.

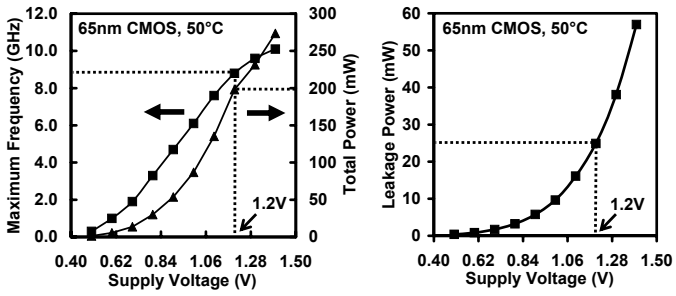


Figure 24.7.5: Maximum frequency, total power, and leakage power measurements versus supply voltage.

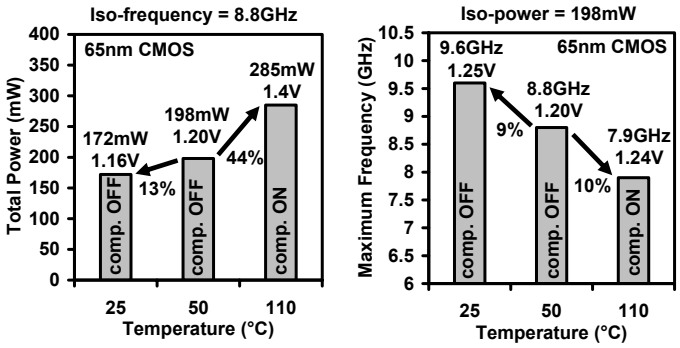
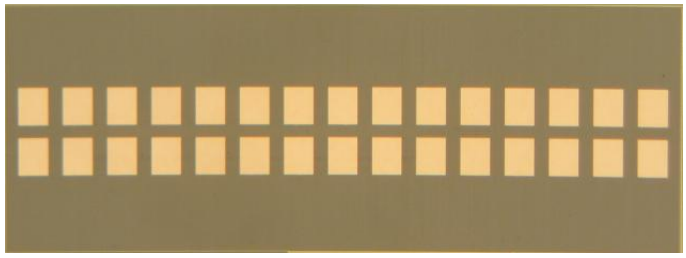


Figure 24.7.6: Total power and F_{max} measurements with simultaneous supply/temperature scaling.



Process	65nm CMOS
Register file layout area	0.017mm²
Worst-case power	198mW at 8.8GHz, 1.2V, 50°C (nominal)
Active leakage power	25mW at 1.2V, 50°C (nominal)
Peak performance	10.1GHz, 273mW at 1.4V, 50°C
Low-voltage mode performance	300MHz, 1.3mW at 0.5V, 50°C
Iso-frequency low-temperature mode	172mW at 8.8GHz, 1.16V, 25°C
Iso-power low-temperature mode	9.6GHz, 198mW at 1.25V, 25°C

Figure 24.7.7: Register file die micrograph and measured performance summary.

